The American Journal of Human Genetics, Volume *89*
Supplemental Data

Abundant Pleiotropy in Human Complex

Diseases and Traits

Shanya Sivakumaran, Felix Agakov, Evropi Theodoratou, James G. Prendergast, Lina Zgaga, Teri Manolio, Igor Rudan, Paul McKeigue, Jim F. Wilson, and Harry Campbell

Supplemental Methods

Independent model

N is the number of genes in the gene pool (*N*=1,380 in this example), *na* and *nb* represent the number of genes associated to trait 1 and trait 2 (in the example this is *na*=8 and *nb*=11). A simple R code for calculating probabilities of exactly *i* overlaps, pp[i] or more then *i* overlaps ppgtr[i]:

```
na <- 8
nb <- 11
N <- 1380


# allocating space
lp <- 0*(1:20);      # log probabilities of exactly i overlaps
pp <-lp;             # probabilities of exactly i overlaps
ppgtr <-lp           # probabilities of over i overlaps


# computing the probability of no overlaps
pp0 <- exp(N*log(1-na*nb/N^2))


# computing the probability of exactly i overlaps and more then i overlaps

for (i in c(1:20))
{
 lp[i] <- i*log((na*nb/(N^2)))  + (N-i)*log(1-na*nb/N^2);
 lp[i] <- lp[i] - lfactorial(i)+ sum(log(N:(N-i+1)));

 pp[i] <-exp(lp[i]);                          # prob of exactly i matches
 ppgtr[i] <- 1 - pp0 - sum(pp[1:i]);          # prob of over i matches
}
```

Degree of surprise criterion

The similarity statistic of the *DS* criterion:

Equation 1

$$I_{adj}(\overrightarrow{x_{AB}}) = -(\sum_{i=1}^{n_g} I(x_i^A = x_i^B = 1)\log p_i^{11} + I(x_i^A = x_i^B = 0)\log p_i^{00} - I(x_i^A \neq x_i^B)\log p_i^{01})$$

where ***I***'s are indicator operators, $x_i^A \in 0,1$is a flag showing whether gene *i* is associated with trait A, p_i^{00}, p_i^{01} and p_i^{11} are parameters of the categorical distribution corresponding to the simultaneously absent, mismatching, and matching genes at location *i* in any randomly chosen disease pair, and n_g is the number of genes in the catalog. *DS* ranks all pairs according to I_{adj}, computes the p-value of each disease pair from the empirical cumulative distribution function of I_{adj}, and identifies pairs of traits at the right tail of the empirical distribution.

Table S1. Differences between the Size of Pleiotropic and Nonpleiotropic Genes when Using the Three Gene Annotation Methods

Status	**Gene size: median (IQR)**			
	Author annotated method (original)	**LD method**	**NHGRI "mapped gene" method**	**Method based on taking all genes in the LD block**
Pleiotropic genes	**45.7kb (114.6kb)**	51.7kb (119.7kb)	76.5kb (180.7kb)	15.6 (57.5)
Non-pleiotropic genes	**38.7kb (94.7kb)**	24.5kb (84.1kb)	58.2kb (127.37kb)	12.0 (53.92)
Pleiotropic vs. Non-pleiotropic (p-value)	**0.072**			
	Gene size: mean (SD)			
	Author annotated method (original)	**LD method**	**NHGRI "mapped gene" method**	
Pleiotropic genes	**131.2kb (261.9kb)**	173.0kb (516.7kb)	163.3kb (269.3kb)	95.3 (345.8)
Non-pleiotropic genes	**95.0kb (166.3kb)**	81.7kb (157.7kb)	135.4kb (226.4kb)	61.4 (147.9)
Pleiotropic vs. Non-pleiotropic (p-value)	**0.007**			

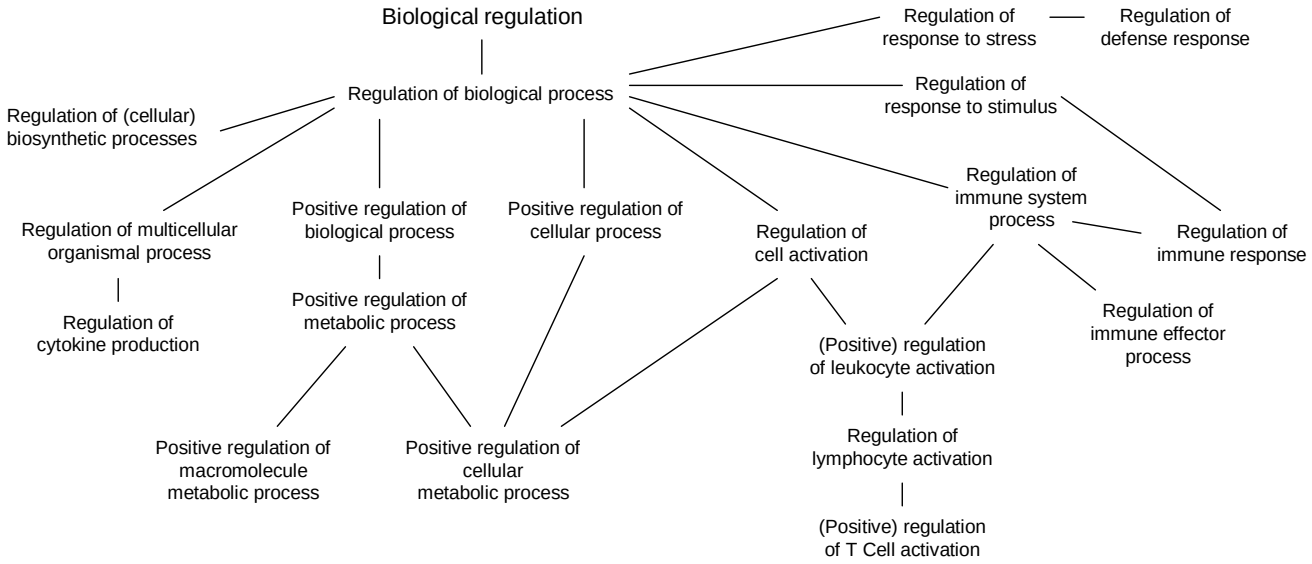


Figure S1. Enrichment of Terms Stemming from “Biological Regulation” among Pleiotropic Genes

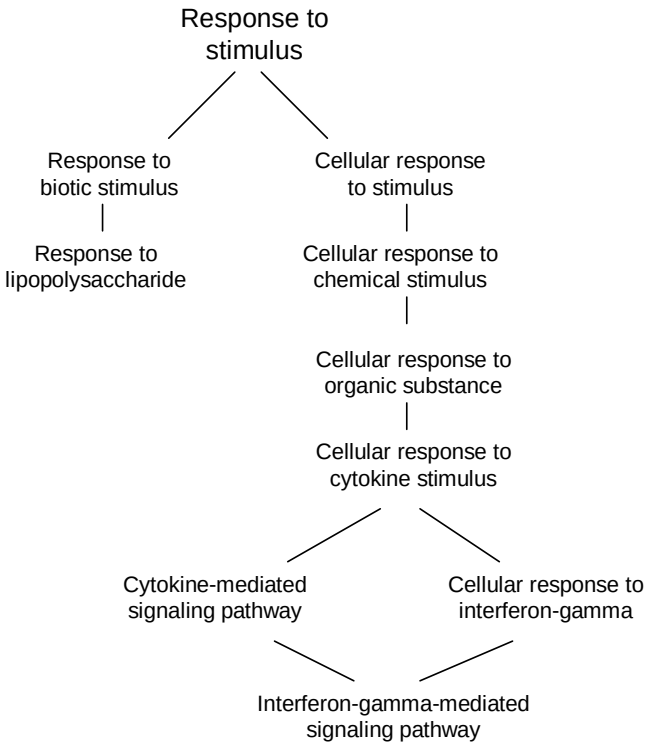


Figure S2. Enrichment of Terms Stemming from “Response to Stimulus” among Pleiotropic Genes

NS = Non-significant

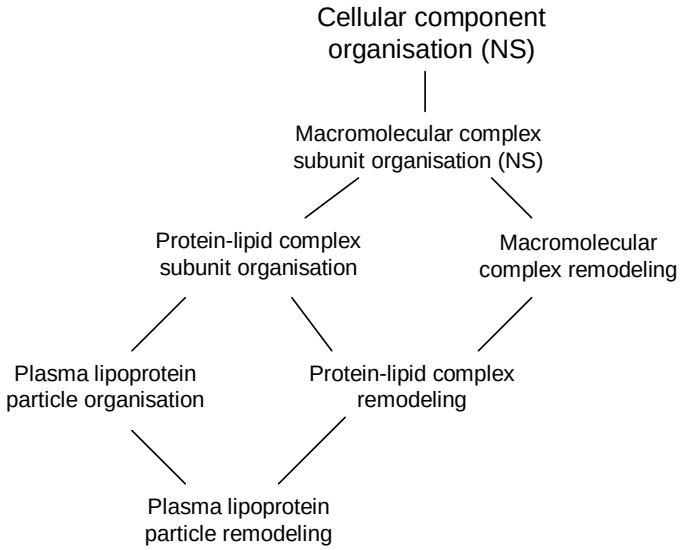


Figure S3a. Enrichment of Terms Related to Lipids among Pleiotropic Genes

NS = Non-significant

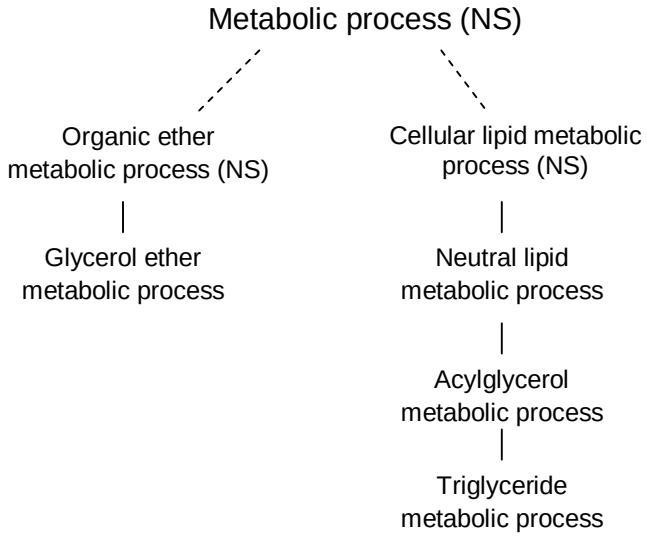


Figure S3b. Enrichment of Terms Related to Lipids among Pleiotropic Genes

NS = Non-significant

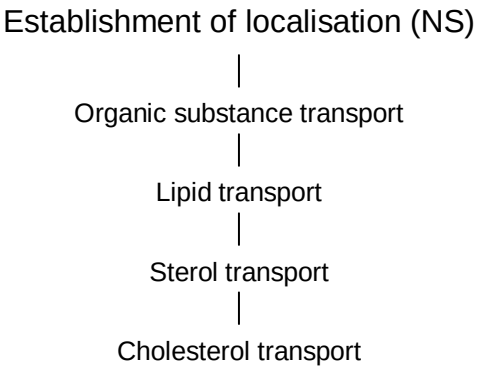


Figure S4a. Enrichment of Terms Related to Lipids among Pleiotropic Genes Associated with Nonimmune Mediated Phenotypes

NS = Non-significant

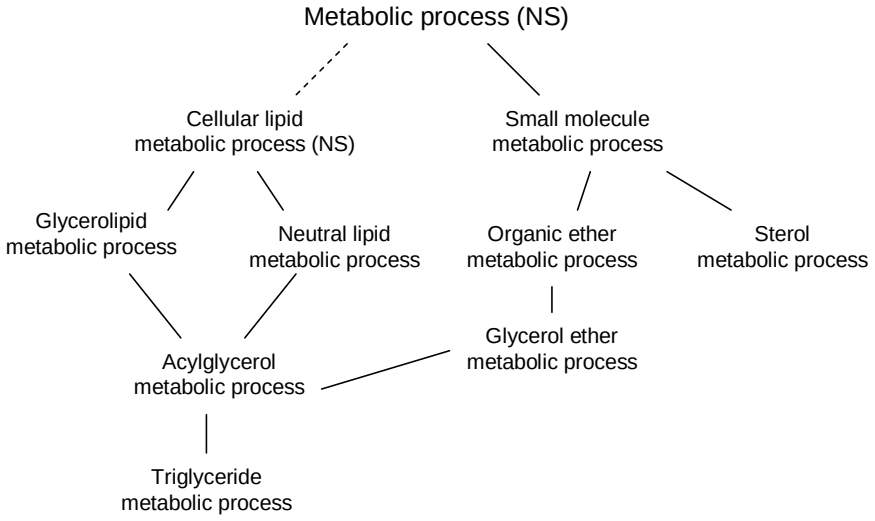


Figure S4b. Enrichment of Terms Related to Lipids among Pleiotropic Genes Associated with Nonimmune Mediated Phenotypes

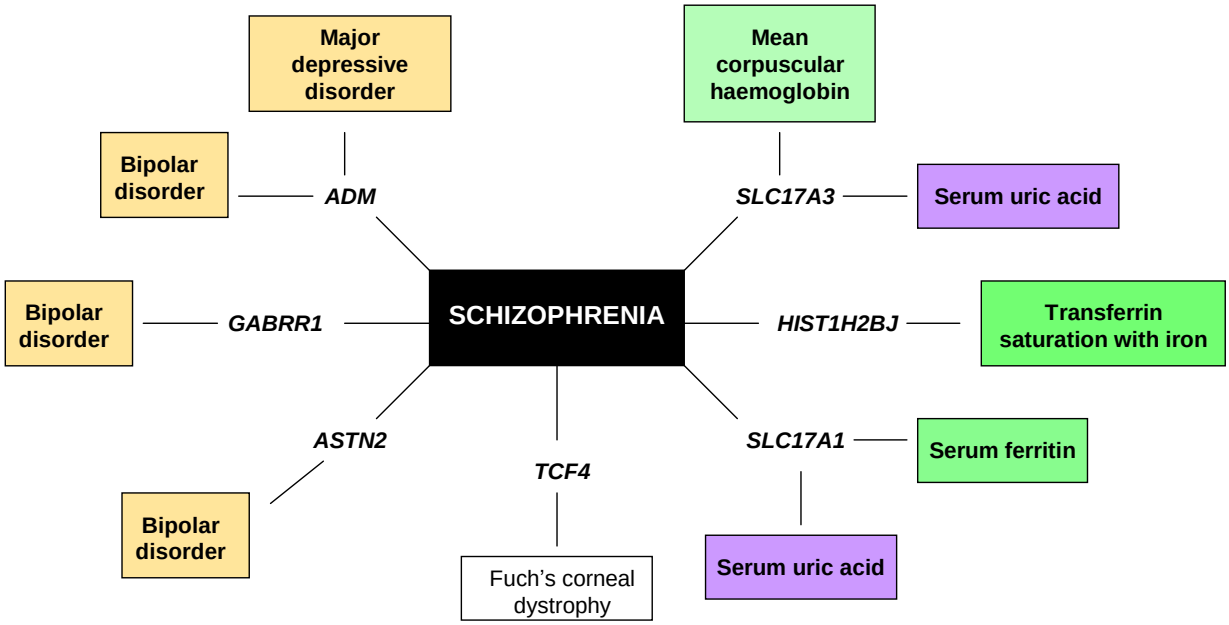


Figure S5. Genes Harboring Variants Associated with Schizophrenia and Other Phenotypes

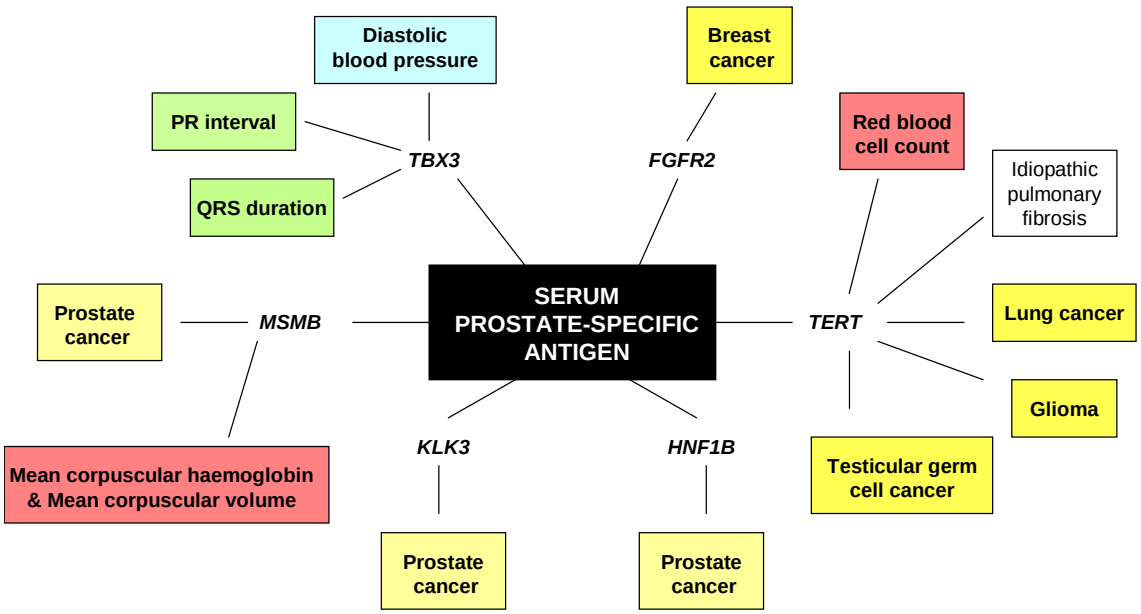
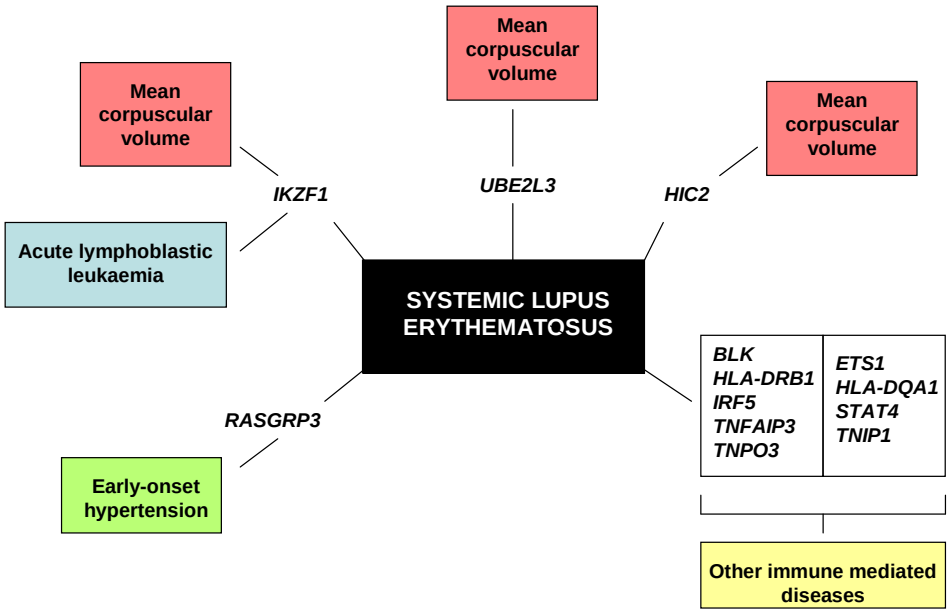


Figure S6. Genes Harboring Variants Associated with Serum Prostate-Specific Antigen and Other Phenotypes



BLK: Rheumatoid arthritis
ETS1: Celiac disease
HLA-DQA1: Celiac disease, Inflammatory bowel disease, Multiple sclerosis, Rheumatoid arthritis, Ulcerative colitis, Vitiligo
HLA-DRB1: Juvenile idiopathic arthritis, Multiple sclerosis, Rheumatoid arthritis, Type 1 Diabetes, Ulcerative colitis, Immunoglobulin A
IRF5: Primary biliary cirrhosis, Rheumatoid arthritis, Systemic sclerosis
STAT4: Systemic sclerosis
TNFAIP3: Celiac disease, Psoriasis, Rheumatoid arthritis
TNIP1: Psoriasis
TNPO3: Primary biliary cirrhosis

Figure S7. Genes Harboring Variants Associated with Systemic Lupus Erythematosus and Other Phenotypes

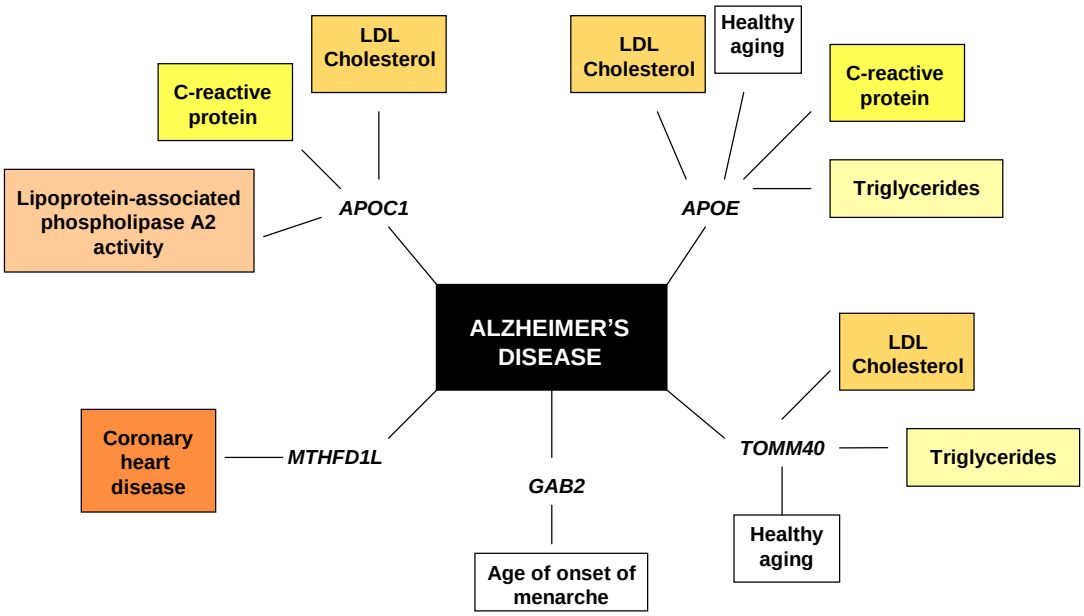


Figure S8. Genes Harboring Variants Associated with Alzheimer's Disease and Other Phenotypes

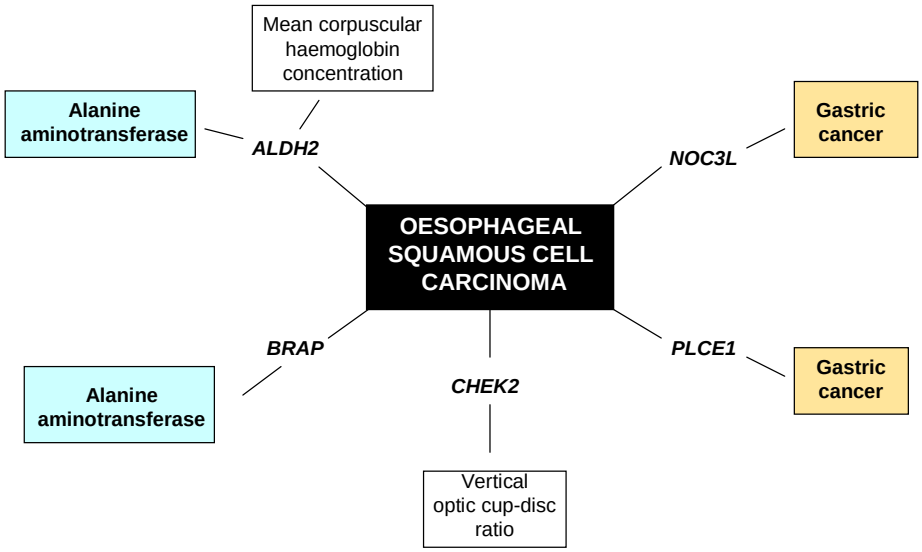
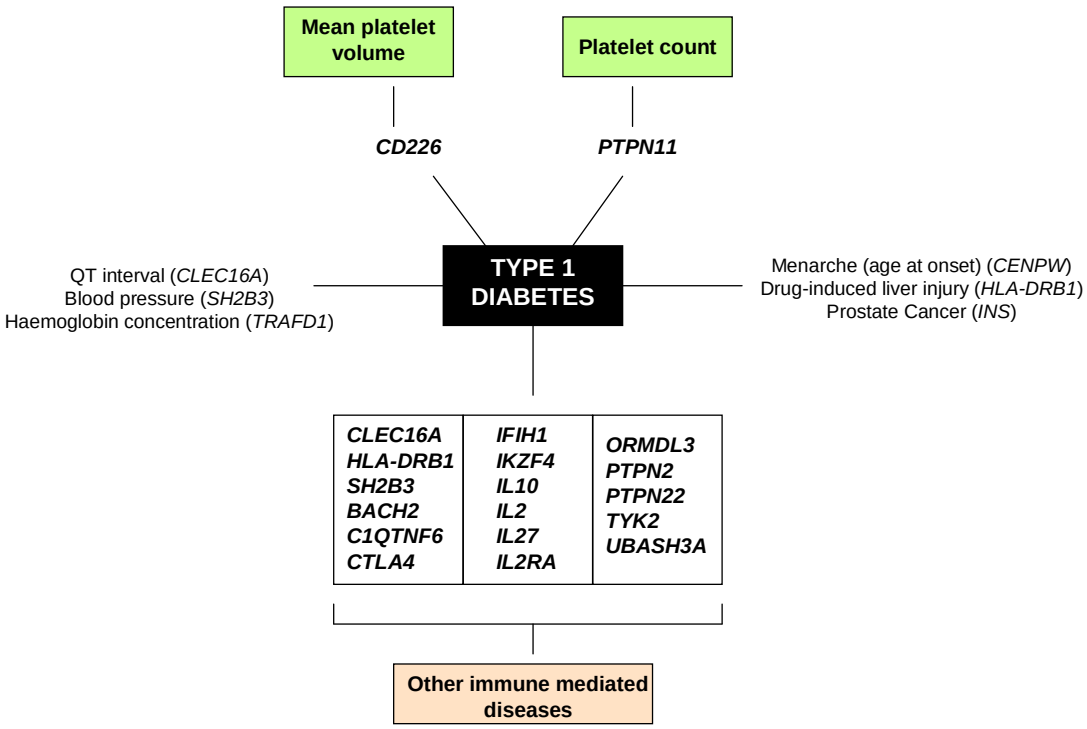


Figure S9. Genes Harboring Variants Associated with Oesophageal Squamous Cell Carcinoma and Other Phenotypes



Genes associated with other immune medated diseases

CLEC16A – Coeliac disease, QT interval
BACH2 – Coeliac disease, Crohn's disease
C1QTNF6 – Vitiligo
CTLA4 – Alopecia areata, Coeliac disease, Rheumatoid arthritis
HLA-DRB1 –Drug-induced liver injury (lumiracoxib), Immunoglobulin A, Juvenile idiopathic arthritis, Multiple sclerosis, Rheumatoid arthritis, Systemic lupus erythematosus, Ulcerative colitis
IFIH1 – Immunoglobulin A, Psoriasis
IKZF4 – Alopecia areata
IL10 – Behcet's disease, Crohn's disease, Ulcerative colitis
IL2 – Alopecia areata, Coeliac disease
IL27 – Crohn's disease, Inflammatory bowel disease (early onset)
IL2RA – Alopecia areata, Crohn's disease, Multiple sclerosis, Rheumatoid arthritis, Vitiligo
ORMDL3 – Asthma, Crohn's disease, Primary biliary cirrhosis, Ulcerative colitis, White blood cell count
PTPN2 – Coeliac disease, Crohn's disease
PTPN22 – Crohn's disease, Rheumatoid arthritis
SH2B3 – Blood pressure (systolic & diastolic), Coeliac disease, Eosinophil count, Haematocrit
TYK2 – Crohn's disease, Psoriasis
UBASH3A – Vitiligo

Figure S10. Genes Harboring Variants Associated with Type 1 Diabetes and Other Phenotypes

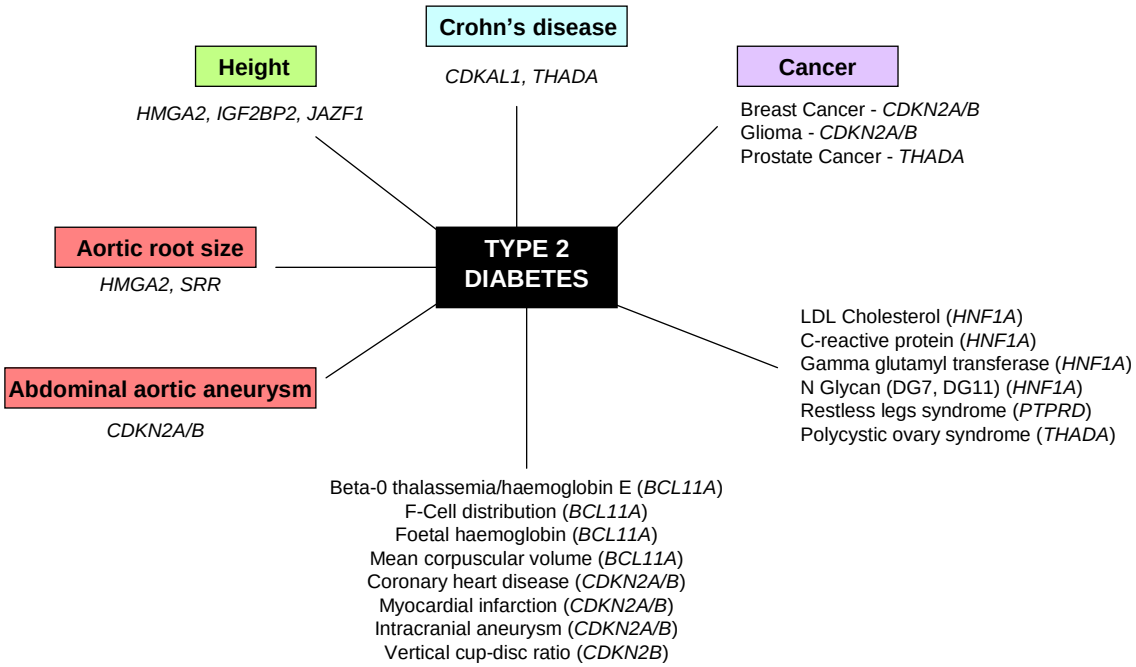


Figure S11. Genes Harboring Variants Associated with Type 2 Diabetes and Other Phenotypes

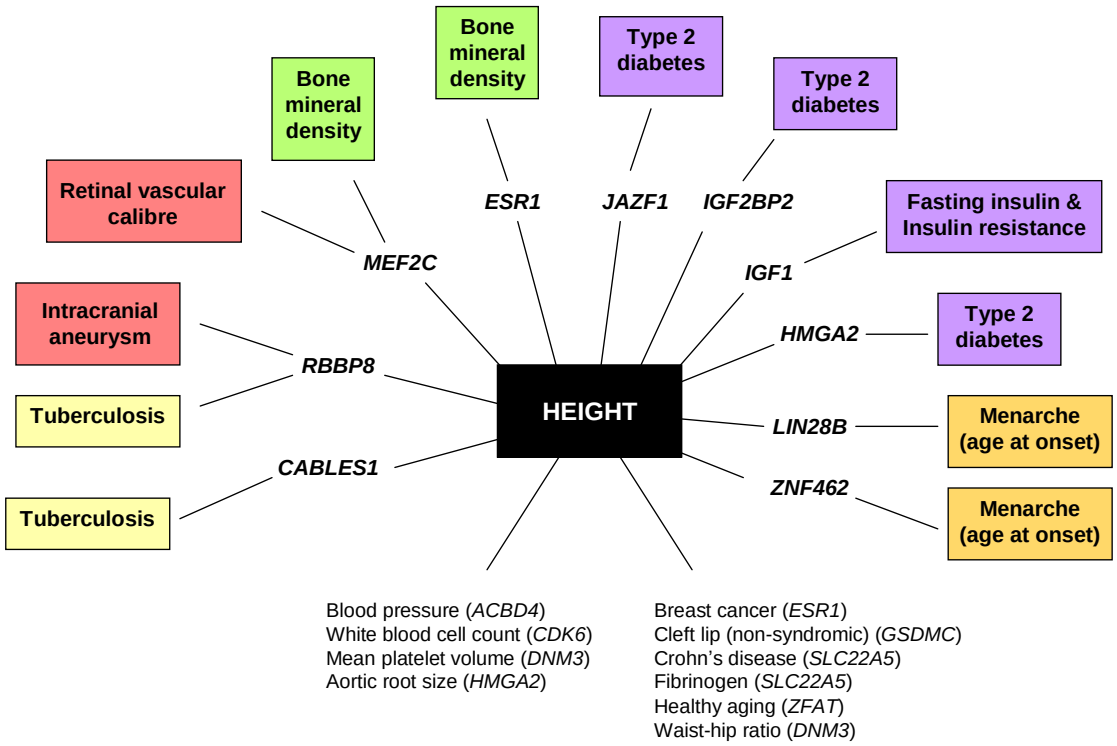


Figure S12. Genes Harboring Variants Associated with Height and Other Phenotypes

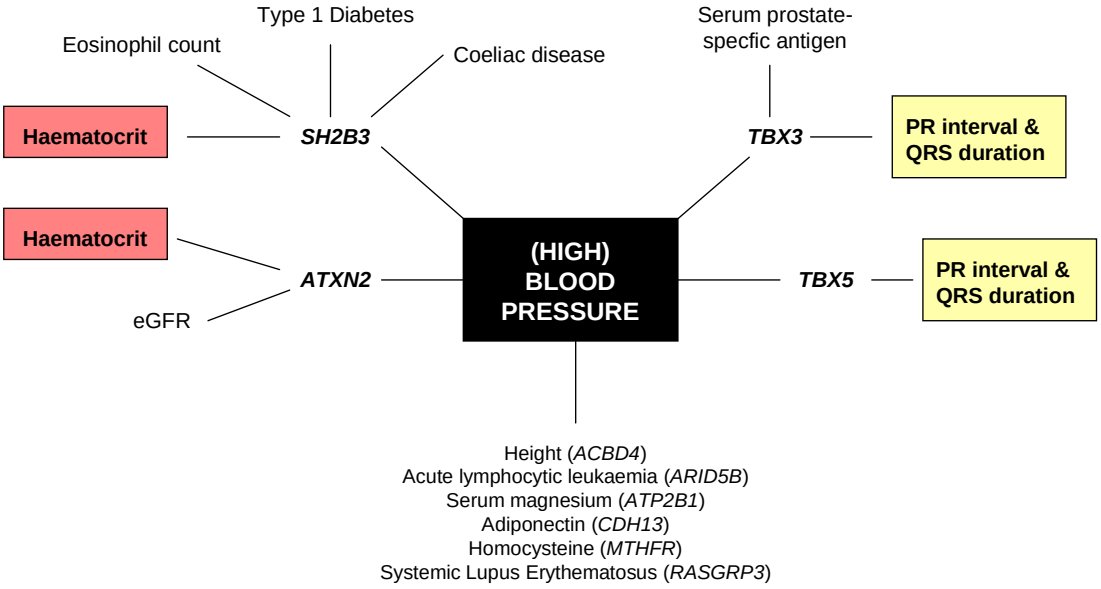


Figure S13. Genes Harboring Variants Associated with Blood Pressure or Hypertension and Other Phenotypes

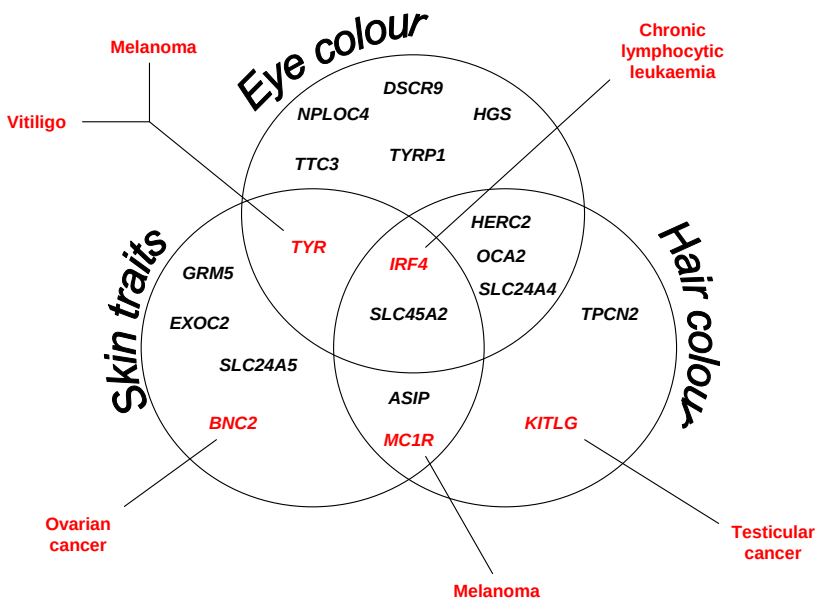


Figure S14. Genes Harboring Variants Associated with Pigmentation Traits and Other Phenotypes
We have not included any study designs other than GWAS, and will underestimate pleiotropy. This situation manifests in this figure, where the known association of *OCA2* with skin pigmentation traits is not accounted for.

References

1. McEliece RJ. 1977. *The Theory of Information and Coding.*Addison-Wesley.

2. Peirce B (1852). Criterion for the Rejection of Doubtful Observations. *Astronomical Journal II.* 45.

3. Peirce B (1877). On Peirce's criterion. *Proceedings of the American Academy of Arts and Sciences.* 13.

4. Bishop, C. M. Novelty detection and Neural Network validation. 217-222. 1994. Proceedings of the IEE Conference on Vision, Image and Signal Processing.

5. Gao, J., Cheng, H., and Tan, P.-N. Semi-supervised outlier detection. 2006. Proceedings of the SAC '06 ACM symposium on Applied computing.

6. Markou M and Singh S (2003). Novelty detection: A review, part 1: Statistical approaches. *Signal Processing.* 83, 2481-2497.

7. Rousseeuw P, Leroy A. 1996. *Robust Regression and Outlier Detection.* 3rd ed. John Wiley & Sons.

8. Tarassenko, L. Novelty detection for the identification of masses in mammograms. 4, 442-447. 1995. Proceedings of the 4th IEE International Conference on Artificial Neural Networks.

9. Chandola V, Banerjee A, and Kumar V (2009). Anomaly Detection : A Survey. *ACM Computing Surveys.* 41.

10. Bishop CM. 2006. *Pattern Recognition and Machine Learning.* Springer.